Summary report on the numeracy reasoning pre-tests taken in May 2013

November 2013

Digital ISBN 978 1 4734 0450 2

WG20177

Contents

1.	Introduction	1
2.	Background to the materials under development	1
3.	Outline of process	1
4.	The pre-test	1
5.	The sample	2
6.	Format of the tests	3
7.	The outcomes of the pre-test	6
8.	Recommendations	9

1. Introduction

This report summarises the findings of the pre-test of numeracy reasoning tests, trialled in May 2013.

The materials were written and developed by Acumina, who also marked the tests. NFER organised the administration of the tests in schools, data-captured the results and provided the statistical analysis. This report has been written by Acumina to give the Welsh Government the information required for future development.

2. Background to the materials under development

Estyn defines numeracy as 'the ability to apply simple numerical facts, skills and reasoning to real-life problems'.[1] The procedural tests assess the ability to recall numerical facts. The reasoning tests focus on the ability to apply what has been taught in a range of contexts that focus on problem-solving. All items assess numerical understanding.

Two reasoning tests, A and B, were developed for each of the year groups 2 to 9. All A and B tests consisted of one stimulus item, presented to the class via Powerpoint by the NFER-employed test administrator immediately before the test, followed by about five written questions in an eight-page booklet. Test administrators were provided with the relevant information on a CD with an accompanying script for use in the classroom, and were asked to familiarise themselves with the materials before delivery.

An additional two reasoning tests, anchor 1 and anchor 2, were developed for each year group. These consisted of questions that could:

- provide a pool of items for inclusion in the anchor tests
- offer replacement items for tests A and B
- be used in the future development of reasoning tests C and D.

All items within tests A and B had already been small-scale trialled, albeit with English schools which were known to encourage their pupils to think mathematically. This enabled us to review and amend items to ensure accessibility and clarity. Because of a lack of time, very few items in the anchor tests had previously been trialled.

3. Outline of the process

The tests were trialled with learners within years 2 to years 9 in schools recruited by NFER. As the timescale did not allow for translation into Welsh, only English medium schools were used. All materials were taken to and from schools by NFER administrators, who ensured security and returned all the papers, used or unused, to NFER.

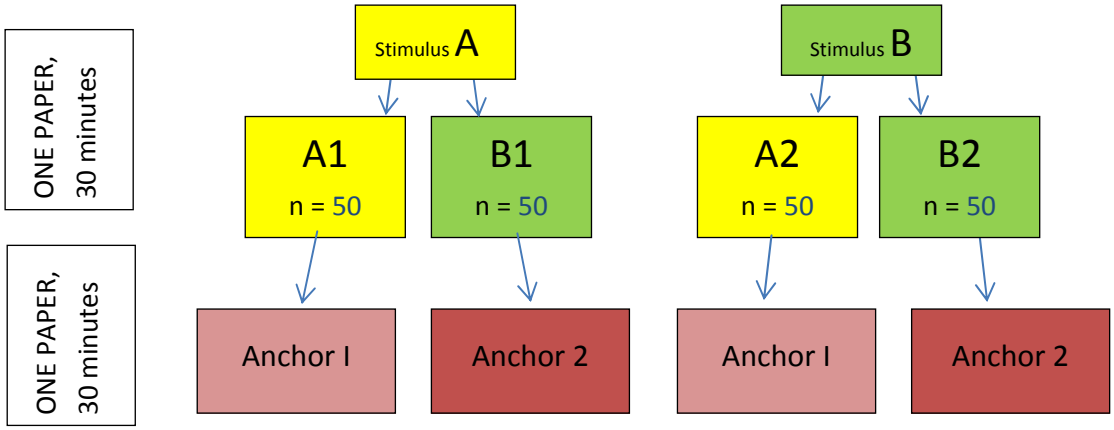
The papers were delivered to Acumina to be marked. Following the marking process, the results were data-captured and analysed by NFER. The analysis has formed the basis of this report and will be used to inform future development.

4. The pre-test

For this trial, four papers per year group were trialled.

[1] Numeracy in key stages 2 and 3, a baseline study. Estyn, June 2013.

In order to discourage copying, and to ensure that learners from as many schools as possible saw the items, a crossover design was used for the pre-test, as shown below.



Most, but not all, teachers provided information about the learners' ability levels, supporting our understanding of item difficulty.

Test administrators were instructed to allow exactly 30 minutes per paper. Most administrators and teachers reported that this time was sufficient to complete the papers. However, some students did not use a calculator where appropriate, which created timing issues.

5. The sample

The target sample size was 200 for each year group, whereby each of the four papers should be seen by 100 learners in four or five schools. Each learner took two papers.

The achieved sample sizes are as follows:

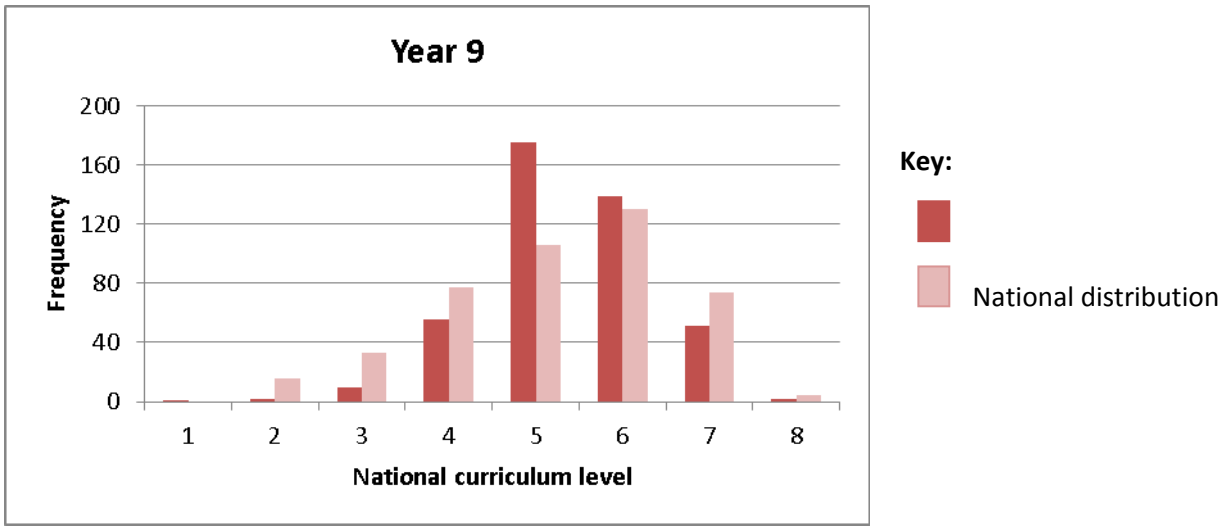
Year group	Test A1	Test B1	Test A2	Test B2	Total for the year
2	52	52	52	53	209
3	52	51	50	49	202
4	56	47	54	47	204
5	52	57	52	60	221
6	60	49	58	50	217
7	53	55	54	53	215
8	54	57	54	58	223
9	57	56	54	54	221

As these tests are so new in style, we anticipated that the results were likely to be disappointing. That schools need to provide greater opportunities for pupils to use numeracy skills, particularly number skills and numerical reasoning, in subjects across the curriculum, is recognised by Estyn[2], and teachers involved in the pre-test told us that their students were unfamiliar with this approach.

[2] Numeracy in key stages 2 and 3, a baseline study. Estyn, June 2013

For that reason, we anticipated excluding the least able students from this pre-test. However, the samples included a significant number of learners working at a level significantly below the level expected for their year group.

For example, as shown in the following graph, the year 9 sample contained a higher number of learners working at levels 2, 3 and 4 than would be expected had the sample reflected the national distribution of levels.



The graph also shows that the number of learners deemed to be working at the highest levels was greater than expected, yet analysis of results suggests that teacher expectation may have been inconsistent. For example, the table below shows the outcomes for one question in the year 9 pre-test; this focused on whether learners could apply their understanding of speed[3]. The question was worth 3 marks in total.

Item name	Mean score	Number of marks gained			Mean score by teacher assessment level			
		1 mark	2 marks	3 marks	NC L4 (n=20)	NC L5 (n=30)	NC L6 (n=34)	NC L7 (n=18)
Robots more walking	32.4%	07.1%	05.3%	26.5%	33.3%	35.6%	22.5%	44.4%

n is the sample size

That the mean score at teacher assessment level 6 is lower than those for levels 4 and 5 may suggest that some teachers have been inconsistent in applying judgements about performance.

6. Format of the tests

6.1 The use of stimulus materials

Each reasoning test A or B starts with a stimulus item that is introduced by the person administering the tests. It is anticipated that this will be someone familiar to the children taking the test as this should reassure learners and reduce test anxiety.

The introduction of stimulus items is an innovative way of assessing numerical understanding: to our knowledge, no country has previously used this approach. We believe it to have many benefits.

[3] Where the number of learners at a given level is very small, the mean score by level is unreliable and has been omitted from the table.

6.1.1 Reducing reading demand

A common criticism of mathematics tests is that, because learners must assimilate printed information, their reading skills as well as their numerical skills are being assessed. Yet it appears to be only in mathematics that there is an expectation that learners should be presented with questions surrounded by minimal text. Problem-solving and the application of numerical skills require contexts, and contexts require words. The use of stimulus materials allows the context and essential information to be presented orally, thus reducing reading demand within the test itself.

6.1.2 Engaging learners

The use of stimulus materials allows complex scenarios to be introduced in an engaging way, capturing interest and motivating learners to give of their best. In some of the trials, younger learners especially were excited about what was being presented and were eager to 'have a go'.

Most teachers said that the stimulus items did indeed engage their learners and helped them settle into the test. However, several teachers indicated that children were disappointed that the other items in the test were not linked to the stimulus item. This was especially noted in years 2 and 3.

6.1.3 Allowing administrators to respond to queries and check understanding

The provision of a script for the stimulus materials ensures that essential information is given to all learners. Those administering the tests are told that they may clarify any aspects of the materials, provided they do not help with the numeracy that is being assessed.

Almost all of the NFER administrators in the trial indicated that the delivery of the stimulus item was straightforward. For example, the table below summarises the responses from the administrators for the stimulus item in year 5 test A1/B1:

	Yes	No
Was the stimulus easy to assimilate?	**5**	**0**
Did it provide all the information you needed?	**5**	**0**
Were there any issues with the script?	**0**	**5**

There was a problem with a video for one item in year 8, and some schools experienced difficulty with an animation in year 3. However, most stimuli were delivered through the medium of Powerpoint which was not problematic.

One teacher noted that his or her learners were taking the test in a hall and it was difficult for some to see the Powerpoint. The materials have been developed for use in a classroom situation, so this will need emphasising to schools preparing for the live tests in summer 2014. The only evidence of a teacher giving inappropriate information during the stimulus presentation was one in year 8 who incorrectly instructed students to amend the key of a graph to show the height of a slide (a slide has variable height).

One teacher responded that it would have been preferable to allow learners to talk together during the stimulus introduction. The instructions will be amended to make it clear that this approach is acceptable.

6.1.4 General feedback for the stimulus items

Most teachers in the pre-test warmly welcomed the use of stimulus materials. For example, the response for the stimulus item in year 2 test A1/A2 was as follows:

	Very poor	Poor	Satis-factory	Good	Very good
Effectiveness - engaging in context			**1**	**2**	**2**
Effectiveness - engaging in test			**2**	**2**	**1**
Interest and enjoyment			**1**	**2**	**2**
Suitability for age group			**1**	**2**	**2**
Ease of presentation				**4**	**1**
Accessibility for SEN			**5**		
Effectiveness in reducing reading demand	**1**			**3**	**1**

	Yes	No
Usefulness, overall	**5**	**0**

In general, the stimulus materials were thought to be engaging and age-appropriate. A few teachers were very enthusiastic, for example:

- 'I would have loved this when I was at school.' (year 5)

However, as for all items in the test, some teachers found them too demanding and almost all teachers expressed concern for learners with special needs.

6.2 The remaining materials

Reasoning tests A and B consisted of one stimulus item, followed by four to six written items, each using a different context. As previously noted, some teachers would have liked to have continued the context used for the stimulus item. Each test included a range of items that aimed to assess the ability to identify processes and connections, represent and communicate, and review. All items were unfamiliar to learners who needed to decide for themselves what to do and how to do it.

Because of the amount of student time taken for the tests, feedback from learners was not sought. Very occasionally children wrote on their test papers, saying, for example, that they were not sure what a word meant or how to do something. One boy, who scored highly, wrote:

- 'This was the best maths ever.' (year 5)

6.3 The allocated time

Generally, teachers and administrators thought that the provision of half an hour per test was appropriate, though some would have liked longer. A few teachers argued that the amount of time was irrelevant as the tests were too demanding and therefore giving longer would not achieve better outcomes.

Schools were encouraged to consider providing year 2 and 3 learners with a break halfway through the test. This was also offered to year 4, at the teacher's discretion. Those that followed this advice responded positively, saying it benefited learners. However, several schools did not allow a break.

7. The outcomes of the pre-test

Mean marks varied but were all disappointing, e.g. 25% (year 3 test A1), 26% (year 4 test A2), 29% (year 5 test B1), 28% (year 7 test B2) and 19% (year 9 test B1).
A multiplicity of factors explains these low scores.

7.1 The unfamiliarity of formal test situations

Statutory testing at key stage 1 was abandoned in Wales in 2002. Following a review of the curriculum and assessment systems in key stages 2 and 3, statutory formal tests for these age groups were removed from 2004/05. Prior to the introduction of reading and numeracy (procedural) tests in 2013, no students in years 2 to 9 had experienced formal tests.

The pre-test provided clear evidence of a lack of understanding of how to work in a test situation. Some learners had clearly copied from each other. Others had little or no knowledge of how to use their time effectively, including checking their answers. As one teacher wrote:

- 'The numeracy was not the problem - children don't have knowledge of how to sit a test.' (year 3)

Many learners left some or most of the questions blank. There appeared to be a reluctance to 'try and see'. Rather, if the route through a problem was not immediately apparent, pupils of all ages appeared reluctant to engage.

7.2 Showing working

Learners of all ages showed difficulty in providing worked solutions. In years 2 to 6, many gave incorrect answers with no supporting working so partial credit was not possible. Or they misunderstood what was required by writing explanations such as 'I know because I did it in my head' or 'I counted on my fingers'. Or they failed to use numerical language by writing mini-essays such as 'Well, first of all I used my 2-times table and I did 2 times 8 and the answer to that is 16 and then I ….'.

Learners in years 7 to 9 appeared more familiar with the need to show working, but few showed the ability to set out work logically, allowing others to understand their approach. Awarding partial credit is difficult if not impossible when numbers appear as if from nowhere.

7.3 Explain questions

Some questions required students to explain their answer. This was found difficult by learners of all age groups. Some teachers recognised the importance of thinking and writing numerically. Others thought this was an unrealistic expectation. For example, when asked if the tests allowed children to show their reasoning ability, one teacher responded:

- 'No, as they needed to explain their answers.' (year 3)

Several teachers asked for the inclusion of Yes / No boxes, or similar, for explain questions. As we believe that drawing conclusions is part of numerical communication, such boxes were not included. A few children omitted their conclusion, but generally we were pleased that most children responded effectively (even if their conclusion was incorrect).

7.4 Mathematical language

Some learners clearly did not know the meaning of words such as 'total', 'fewer than', 'height', 'perimeter', 'area' and 'profit'. This was worse for the younger children but was still evident for older year groups. A common error at years 2 to 5 was to ignore the word 'more'. So, for example, for the simple question 'Alun has £5 more than Jen. Jen has £3. How much does Alun have?' the common error would be that Alun has £5.

The Framework lays out an expectation for relevant numerical vocabulary but at the time of the pre-test many teachers appeared unfamiliar with the requirements, querying their inclusion.

7.5 Reading demand

Reading demand within the tests was minimised by using simple sentence structure (subject-verb-object), through use of diagrams and 'speech bubbles' and by ensuring clear layout of information. However, these tests must, by their very nature, require more reading than the procedural tests since problem-solving requires a context in which to place the numerical demand. In some questions, we trialled use of phrases rather than complete sentences, but some teachers disliked this approach arguing that it showed 'poor literacy'.

The general guidance for the tests allows administrators to read text where appropriate, but it was clear that some teachers in the trial were unaware of this guidance. One teacher in year 3 argued that the usual practice would be to read a whole question then give learners a chance to respond before reading the next item.

There was evidence in all year groups of learners not reading key information, even when it was in bold or bulleted. Some pupils had been taught to highlight important information. These pupils tended to out-perform those who may have skim-read or ignored words on the page.

7.6 Lack of number sense and checking strategies

'Mathsworld' is a term used by the mathematical community to describe the inability to use number sense when working on mathematical problems. It is as if common sense is left behind so that, for example, learners state confidently that someone must be 180 years old rather than re-visiting their method.

Throughout these tests, at all year groups, some children displayed a worrying lack of number sense giving answers that made no sense within the given context. Many learners adopted an 'if in doubt, add' strategy to a wide range of problems. There was almost no evidence of checking strategies.

7.7 Use of a calculator / inefficient methods

The procedural tests assess whether learners can process calculations, therefore no calculators are allowed. To avoid learners spending time on complex calculations, in the reasoning tests calculators should be available from year 5 upwards.

Several teachers, especially in years 5 and 6, queried the availability of calculators, stating that their school policy was not to use them. Even in year 8, one school responded that their school did not allow students in that year group to use a calculator. These learners were clearly disadvantaged as they spent an inappropriate amount of time trying to work things out by hand. Long multiplication and division were used inappropriately (and mostly poorly) for very large numbers, and some learners filled the whole page with repeated addition or subtraction.

Many learners appeared not to know how to use a calculator effectively. In all year groups, some learners used their calculator inappropriately, for example adding £5.99 and 85p to give the answer £90.99 and many struggled to interpret, for example, 1.5 in the context of money, giving the answer £1 and 5 pence.

Estyn[4] has recently reported concerns that 'in key stage 3, too many pupils use calculators for simple calculations, where a mental or written method would be more appropriate'. However, in this pre-test, the converse was found as many learners in all year groups failed to show their reasoning capability because they focused on operating procedurally.

7.8 Problem-solving

Problem-solving in real life requires more than one step towards a solution, and this was true of the majority of items in the pre-test. Teachers of all year groups queried this approach. For example:

- 'There were too many steps required.' (year 2)
- 'Not enough scaffolding.' (year 4)
- 'Needs to be one-stage questions.' (year 6)
- 'Too many aspects to consider.' (year 7)
- 'More structure needed to enable students to start.' (year 8)

Some teachers suggested that the items would be better in the form of first do this, then do this, and so on. However, this would remove the reasoning requirement.

It is clear that many learners are not used to identifying and selecting the numerical skills required then proceeding logically towards a solution. As such, it is not surprising that the outcomes of the pre-test were mostly disappointing. However, it was encouraging to read the comments of some teachers who clearly understood the philosophy behind the development. For example:

- 'Questions provided opportunities for pupils to show their learning.' (year 3)
- 'Challenging questions encouraging excellent thinking skills.' (year 4)
- 'Good opportunities to explain their thinking strategies for each question.' (year 5)
- 'Reasoning is difficult but important. Problem-solving is often an area that pupils find hard, but this needs to be addressed.' (year 6)
- 'Good thinking questions, with lots of maths involved.' (year 9)

Learners who are used to being told what to do and how to do it are unlikely to perform well on items that require them to think and act independently. As Estyn[5] reported:

- 'Pupils … are reluctant to apply (their skills) to solve problems, particularly in the context of other subjects such as science and technology.'

[4] Numeracy in key stages 2 and 3, a baseline study. Estyn, June 2013.
[5] The Annual Report of Her Majesty's Chief Inspector of Education and Training in Wales, 2011-2012

8. Recommendations

We recognise that the outcomes from the pre-test must be treated with a degree of caution since the sample sizes are small and not always representative of national performance. Nonetheless, there is a clear issue in terms of level of performance.

As noted earlier, these tests are unique in terms of focusing solely on reasoning capability, so it is a learning curve for us as test developers in terms of what can reasonably be expected. We recognise that we have sometimes been over-ambitious in terms of the questions given for a particular year group. However, our understanding is informed by the performance of reasoning items within key stages 1, 2 and 3 in Wales and England[6]. The challenge is to create tests that reflect the framework, encouraging teachers and learners to be ambitious, whilst also being realistic.

As these tests will be new to schools in May 2014, it is likely that performance will be considerably lower than ideal as teachers and learners begin to understand the new framework and the expectations regarding numerical reasoning. However, the fact that a very few learners scored highly on the pre-test offers reassurance high standards are attainable.

Our recommendations are therefore as follows:

8.1 Format of the tests for years 2 and 3

Teacher and administrator feedback and evidence from the marking process suggest that the reading demand in years 2 and 3 is problematic. There is also evidence that these learners struggled to move through a range of different contexts.

Recommendation 1:

- For years 2 and 3 only, introduce two stimulus items per test (15 minutes each) with a small number of linked items for each. (As most of these contexts will be new, a further pre-test will be needed for these year groups to ensure the accessibility and effectiveness of the items.) Ensure schools deliver these two sets of materials at different times to avoid both test fatigue and confusion between the two sets of materials.

8.2 Format of the tests for years 4 to 9

The introduction of stimulus materials was generally very well received. No issues were reported in terms of delivery of the materials via Powerpoint though some schools experienced difficulties with animations that had been created especially for the pre-test. This issue will be investigated further.

Recommendation 2:

- For years 4 to 9, keep the format of the tests as used within the pre-test, i.e. one stimulus item to be introduced before the test with a range of different items to follow. Ensure that schools understand the format well in advance of the tests being taken, and that the stimulus is designed to be delivered within a classroom environment rather than a hall.

[6] The Acumina team includes the project director for the assessment of number skills across curriculum for Wales, 2008 – 2009, who was previously the Project Director for test development, ks2/3, mathematics, for Wales, 2003 – 2007. In addition, Acumina has written and developed items used in national curriculum tests for key stages 2 and 3 in England, and also for single level tests, levels 3 to 8. All materials included items that focused on reasoning.

8.3 Timing of the tests

Teacher and administrator feedback suggests that the allotted test time of half an hour (excluding time to deliver the stimulus material) is reasonable, provided learners understand the importance of using their calculator appropriately.

Recommendation 3:

- Keep the timing at 30 minutes per test, but encourage schools to allow access to a reading book so that those who finish early, and have checked their work, have something to do whilst others continue working. As noted previously, for years 2 and 3, ensure that the 30 minutes is delivered as two distinct sets of 15 minutes. In addition, for year 4 only, allow a break within the test at the teacher's discretion.

8.4 Difficulty level of the tests

It is clear from the outcomes of the pre-test that at least some of the items need to be made more accessible to allow greater opportunities for learners to show what they know and can do.

Recommendation 4:

- Where possible, ease the stimulus item to widen the range of ability that can engage with the materials and progressively increase demand so that questions at the end of the paper challenge the most able and give clear messages about expected level of performance against the new framework.

8.5 Learners with additional learning needs

Almost all teachers and administrators indicated that learners whose attainment was significantly below the expected level for their year group experienced difficulties in accessing the tests and were sometimes dispirited as a result. This raises issues about what can reasonably be expected by learners with additional learning needs. The pre-test provided clear evidence that some learners had not been taught the relevant mathematics, for example percentages, area or volume, so were unable to access reasoning items that required the application of these skills.

There is a clear tension between over-expectation in what can be achieved and under-expectation in terms of such learners not being exposed to a cognitively demanding curriculum. In the small-scale trials of reasoning items, teachers were surprised at how some learners deemed to be low-ability out-performed others who were thought to be more able. (The converse was also true with learners labelled high-ability who sometimes struggled to apply reasoning skills.)

The UK has one of the widest achievement gaps in the world[7]. For Welsh students ages 14 and 15/16, there is on average a gap of 32 to 34 per cent between what children living in poverty achieve compared to other students[8]. The Welsh Government has designed the framework to be inclusive of all learners. However, it acknowledges that the framework 'describes a continuum of development and learners may progress further or faster in some aspects than in others, with achievements spanning several years[9]'.

[7] Speech by David Laws, Schools' Minister, Department for Education, 5 March 2013 https://www.gov.uk/government/speeches/closing-the-achievement-gap

[8] Poverty and low educational achievement in Wales: Student, family and community interventions, Joseph Rowntree Foundation, February 2013

[9] http://learning.wales.gov.uk/docs/learningwales/publications/130415-lnf-guidance-en.pdf, page 11

One solution is to allow schools to enter some learners for a reasoning test that is linked to the framework for a different year group. This has a clear advantage in that more positive outcomes should be expected. However, the disadvantages are that this approach might reinforce a lack of self-worth and might lower learner and teacher expectations. There is also an issue in terms of standardisation of scores for the year group, and establishing measures of progress.

Recommendation 5:

- That the above should be discussed in detail by the Welsh Government in order that they can provide schools with clear guidelines on the suitability of the reasoning tests for those with additional learning needs.

8.6 Preparation for the tests

Teachers and learners would benefit from being familiar with the style and layout of the test items, understanding, for example, the purpose of working boxes. It is also clear from the feedback on the tests that teachers would benefit from understanding the philosophy behind the introduction of the tests: too many believed that 'numeracy' and 'reasoning' were separate entities.

Recommendation 6:

- If possible, review or replace the sample reasoning materials in the light of the outcomes from the pre-test in order to ensure their suitability for each year group. Additional teacher support material could be produced to complement the sample materials, helping teachers to understand why reasoning in numeracy is important and also how to prepare their learners for the test format.

8.7 Marking of the tests

The pre-test has enabled us to review the markschemes to ensure that they are as accessible as possible. However, the nature of reasoning items is such that learners can choose their own method: this in turn means that awarding partial credit is rarely simple. This is especially true in years 7, 8 and 9 in which the problem-solving nature of the questions is more advanced.

Recommendation 7:

- Implement external marking for years 7, 8 and 9, and ensure that markschemes are as accessible as possible for all other year groups.

8.8 Understanding and building on the outcomes of the tests

The scores from the reasoning tests, including standardised scores and progression scores, will offer teachers good insight into the students' performance. However, we believe that it is equally as important that teachers interpret what they are seeing, for example by recognising common errors, and by seeing a range of responses that illustrate effective and less effective methods.

Recommendation 8:

- Include exemplars and brief commentary within the markschemes to support marker and teacher understanding. This information could be used in a variety of ways.

Acumina, September 2013